Newt-omics: a comprehensive repository for omics data from the newt *Notophthalmus viridescens*

Marc Bruckskotten[1], Mario Looso[1], Richard Reinhardt[2], Thomas Braun[1,*] and Thilo Borchardt[1,*]

[1]Department of Cardiac Development and Remodelling, Max Planck Institute for Heart and Lung Research, Ludwigstrasse 43, D-61231 Bad Nauheim and [2]Max Planck Genome Centre, Carl-von-Linne-Weg 10, 50829 Cologne, Germany

ABSTRACT

Notophthalmus viridescens, **a member of the sala-mander family is an excellent model organism to study regenerative processes due to its unique ability to replace lost appendages and to repair internal organs. Molecular insights into regenerative events have been severely hampered by the lack of genomic, transcriptomic and proteomic data, as well as an appropriate database to store such novel information. Here, we describe 'Newt-omics' (http:// newt-omics.mpi-bn.mpg.de), a database, which enables researchers to locate, retrieve and store data sets dedicated to the molecular characteriza-tion of newts. Newt-omics is a transcript-centred database, based on an Expressed Sequence Tag (EST) data set from the newt, covering ~50 000 Sanger sequenced transcripts and a set of high-density microarray data, generated from regene-rating hearts. Newt-omics also contains a large set of peptides identified by mass spectrometry, which was used to validate 13 810 ESTs as true protein coding. Newt-omics is open to implement additional high-throughput data sets without changing the database structure. Via a user-friendly interface Newt-omics allows access to a huge set of molecu-lar data without the need for prior bioinformatical expertise.**

INTRODUCTION

Comprehensive repositories for the standard model organisms mouse, rat, zebrafish, *Drosophila* and *Xenopus* (1–5)

provide access to all levels of sequence data sets, including genome, transcriptome and proteome data. Detailed information for single genes usually comprises knowledge about intron and exon regions, splicing signals, as well as sequence and functional annotations. To enable a user-friendly handling of such information databases are most often accessible through graphical user interfaces that allow combinatorial searches for different properties of single genes or gene families. However, for non-standard model organisms, very little information from publically accessible data have been collected and organized in a user-friendly form. This situation prevents dissemination of useful research information to a broader research community and keeps such model organisms in splendid isolation. One of these organisms is the red spotted newt *Notophthalmus viridescens* (vertebrata, order caudata, family Salamandridae, genus *Notophthalmus*), known for its exceptional regenerative capabilities for more than 200 years. The newt does not only possess the ability to entirely replace lost appendages (6–8) but also regenerates the lens (9,10), parts of the central nervous system (11,12), and the heart (13–15). These unique features make the newt an excellent model to study fundamental processes of tissue regeneration. Newts have been helpful to study embryonic development of vertebrates (16), which led to the first Nobel prize in developmental biology. However, the introduction of other model organisms with shorter life cycles and better characterized genomes have led to a massive decline in the use of newts for basic scientific research. Moreover, the estimated genome size of the newt is up to 10 times larger than that of humans. These circumstances have severely impeded genome projects despite the increasing speed and capacities of modern sequencing machines and assembly algorithms and slowed down development of

methods devoted to genetic manipulation of this animal. As a result of these drawbacks, approximately 100 non-redundant protein sequences for the newt are available in the NCBInr database, although a set of almost 11 000 sequenced Expressed Sequence Tags (ESTs) from regenerating hearts of the newt *Notophthalmus viridescens* exists (17).

Here, we introduce Newt-omics, the first public available data repository for *N. viridescens*. In addition to EST data sets and a set of high-density microarray data, Newt-omics contains a large set of peptides identified by mass spectrometry (18). Newt-omics provides a user-friendly gateway to large molecular data sets, which helps to make this attractive model organism more accessible even for researchers with limited experience in bioinformatics.

RESULTS

Transcriptome data

To expand our knowledge of transcriptional changes that occur during regeneration and to provide a data set for the Newt-omics database, we generated a normalized cDNA library from various time points after newt heart injury covering the entire cardiac regeneration process. The cDNA library was plated, 100 000 individual clones were picked, amplified and the products spotted to produce custom-made cDNA microarrays.

Complementary DNA samples from nine different time points of regenerating newt hearts plus sham controls were compared with undamaged samples. Biological and technical replicates were pre-processed and normalized. Resulting expression values for each cDNA spot were stored in the Newt-omics database along with statistical significance values and linked to their dedicated transcript. Around 52 000 EST clones were selected for Sanger sequencing based on expression changes that were detected during microarray expression analyses. After filtering, clipping and removal of contaminants, sequences were assembled into 26 594 unique transcripts with an average length of 642 bp.

Annotation of sequence data

BLAST homology searches were performed using several BLAST algorithms and databases. For this purpose, an automated annotation and quality filter pipeline was programmed. The BLAST was performed by BLASTcl3 controlled by a Unix shell script. We used blastn, blastx and tblastx on NCBI's NR (protein), NR (Nucleotide), EST and High Troughput Genomic Sequences (HTGS) databases. All hits were sorted according to their taxa. For each taxon, we performed a quality rating and filtering by regular expressions and checked for keywords providing a robust quality level for a single hit (like 'mRNA','cDNA', 'clone' or 'genomic'). The number of blast hits and their quality were calculated for each taxon independently. In total, we evaluated 90 different taxa. Since we wanted to achieve a maximum quality of annotations per transcript with a minimum number of annotations per taxon, we assumed that an annotation with a NCBI NR entry is of higher quality than an

annotation with an EST entry, which by itself is of higher quality than an annotation by a HTGS entry. Based on this assumption we separated BLAST hits with the highest annotation quality from lower quality hits by computing a quality rank. The rank algorithm was based on a dictionary of keywords (e.g. cDNA, clone, mRNA, predicted) reflecting hits of limited information. Database entries containing one or more such keywords were marked as low quality hits. Our annotation was complemented by hits from lower quality groups according to the number of low quality keywords and the homology score if the minimum number of annotations was not reached for a single taxon within the highly ranked group. Our annotation process collected at least three top hits per taxon, BLAST algorithm and database. Due to this workflow, the number of selected Blast hits that were required to annotate a transcript sufficiently decreased with the extent of sequence similarities found. Interestingly, we detected a substantial number of sequenced and assembled transcripts, which did not share any reliable sequence similarity to database entries. We concluded that these sequences were either unique for urodele amphibians or had not been identified in other species yet.

Functional annotation

Subsequent to transcript annotation, we assigned transcripts to different functional groups. Since the extent of Gene Ontology (GO) annotations for amphibians is limited (19) we performed separate BLAST searches against Uniprot databases (*e*-value threshold $< e$-20) from mouse, human, zebrafish, chicken and cow. The mammalian organisms human, mouse and cow have well annotated sequence information. The chicken is the closest relative with a substantial number of GO annotated proteins in the evolutionary tree to newts. The zebrafish served as a model organism for tissue regeneration, although the annotation is comparatively poor. We only took the best-rated similarity hit per taxon with at least one GO annotation to avoid redundant assignments. GO annotations for each transcript were stored in a taxon-dependent manner in the Newt-omics database. We also assigned known functional domains, interaction partners, protein families and pathways to individual transcripts based on Uniprot entries.

Proteome data

To obtain a comprehensive set of proteins that are present during organ regeneration, we performed mass spectrometry (reverse-phase nano-LC-MS/MS) experiments on several newt heart and tail samples isolated at different time points of regeneration. We generated stable isotope labelling with amino acids in cell culture (SILAC) labelled proteins *in vivo* (20) to facilitate quantification and to compare protein levels between different biological samples with pulsed SILAC (21). Annotated ESTs from the transcriptome analysis were not used directly for mascot searches due to a number of inherent restrictions [discussed in detail in Ref. (18)]. Instead, all newt ESTs were reverse translated into three reading frames to generate a MASCOT-conform search database. Linkage of peptides

to ESTs was achieved by computing the absolute position of the peptide within a corresponding EST. The maximum false discovery rate was set <1% for peptide and protein identifications using decoy target databases (22). We could experimentally validate 15 169 different peptides corresponding to 13 810 ESTs that were stored in the Newt-omics database.

Design of the database

The principal aim of the Newt-omics database is to provide a comprehensive and dynamic repository for 'omics' data from the newt *N. viridescens*. We intended to integrate different sets of data obtained from sequencing approaches, quantitative transcriptome or proteome studies, as well as bioinformatics analysis. The Newt-omics design is based on an Entity Relationship Model (ERM) (23) that connects data sets by biological context and avoids compartmentalization. Hence, we prioritized the selection of biological entities (such as EST, transcript and peptides) over experimental entities (such as regeneration stages of the heart). Experimental settings were generated as attributes of single entities, that allowed direct linkage between biological entities independent of their experimental origin. Entities such as EST, transcript and peptide were defined as core tables and additional entities such as transcript annotation, functional annotation of transcripts and ORF of transcripts as section tables. The central object of our database design is the transcript. Transcripts mediate queries across all sections of the database, which is the prerequisite to incorporate an advanced Graphical User Interface (GUI) search engine. A transcript table connects the main sections annotation, functional annotation, expression and peptide (Supplementary Figure S1). The database can be upgraded by implementation of future experiments. For example, new sequencing data from next generation sequencing approaches can be included into the database by adding single sequences to the EST table and by adding results of an assembly as transcripts to the transcript table. Likewise, corresponding annotations can be appended to the annotation tables. The annotation section instantiates a defined transcript and represents all annotation results of a homology search at a specific date. Furthermore, identifications of new peptides can be directly mapped to corresponding transcripts. This approach allows us to maintain consistency of old identifiers and statistics. Linkage of peptides to transcripts was realized by a mapping table, which translates ESTs into all possible open reading frames (ORF). Only experimentally verified peptide sequences that map to an ORF table are included in the database.

Individual sections can be updated separately since, e.g. annotations or ontologies usually change more often than transcript sequences or associated peptide data. The update routine for the annotation section allows different instances as discussed above. Since our functional annotations depend on similarities to a Uniprot protein, features describing a transcript can be extended by any other information appended to the Uniprot identifiers. Our database design allowed us to implement a database frontend with the central object 'transcript'. Starting from

a transcript in Newt-omics, all affiliated data can be directly accessed.

Graphical user interface

To facilitate easy access to the database we developed a web-based user interface. The user interface was designed using PHP/HTML/JS scripts. Database searches can be initiated from the four main windows 'Transcripts', 'Blast', 'Peptides' and 'Expression' (Figure 1A). The main window 'Transcripts' contains: (i) a transcript-centred search enabling access to internal EST or transcript identifiers, transcript length and name; (ii) an annotation-centred search allowing selection of the annotation source (algorithm and database) and searches for keywords or significance (*e*-value); (iii) a functional annotation search allowing searches for GO term numbers, Uniprot Identifiers, pathways, protein families and Interpro domains for single transcripts or a group of transcripts of interest. Complex searches can be performed by a combination of logical AND or OR statements within the three search tabs (Figure 1A). Searches can be filtered by expression changes and by transcripts for which at least one peptide has been experimentally validated. Search results from the 'Transcripts' window are displayed in a table, listing transcripts that match to a query (Figure 1B). This results table can be sorted by transcript ID and number, the number of ESTs corresponding to the transcript, the length of the assembled transcript, the number of annotations from a BLAST search and the number of hits in functional annotations. A detailed list, showing all individual annotation results, can be expanded from a tab on the bottom of the results page. Here, the blast alignment can be visualized by moving over the alignment link of an annotation hit. The main window 'BLAST' enables similarity searches to any sequence entry in the Newt-omics database using the BLAST tools. Searches in this window can be filtered by *e*-value, BLAST algorithm and by data source (Figure 1A). Results of a BLAST search are displayed according to NCBI BLAST results, with a graphical coverage, a results list providing direct links to matching transcripts and a sequence alignment view (Figure 1B). The main window 'Peptides' permits searches for experimentally validated peptide properties. The form has a query field for core attributes of peptides such as, length, mass, mascot score and modification. These core attributes can be chosen from individual experiments sets or RAW files (Figure 1A). The results table of a peptide search displays matching ESTs, corresponding peptide sequences and peptide properties, including length of the peptide, modification and molecular mass (Figure 1B). The main window 'Expression' provides access to microarray expression data from regenerating newt heart. The search form is able to combine queries for fold changes and *P*-values for each individual time point of the experiment (Figure 1A). Search results are visualized in a heat map, displaying all matching transcripts with their ID and name and expression values with corresponding *P*-values for each time point of an experiment (Figure 1B).

Once a transcript of interest has been identified by any of the four main search windows, its selection directly
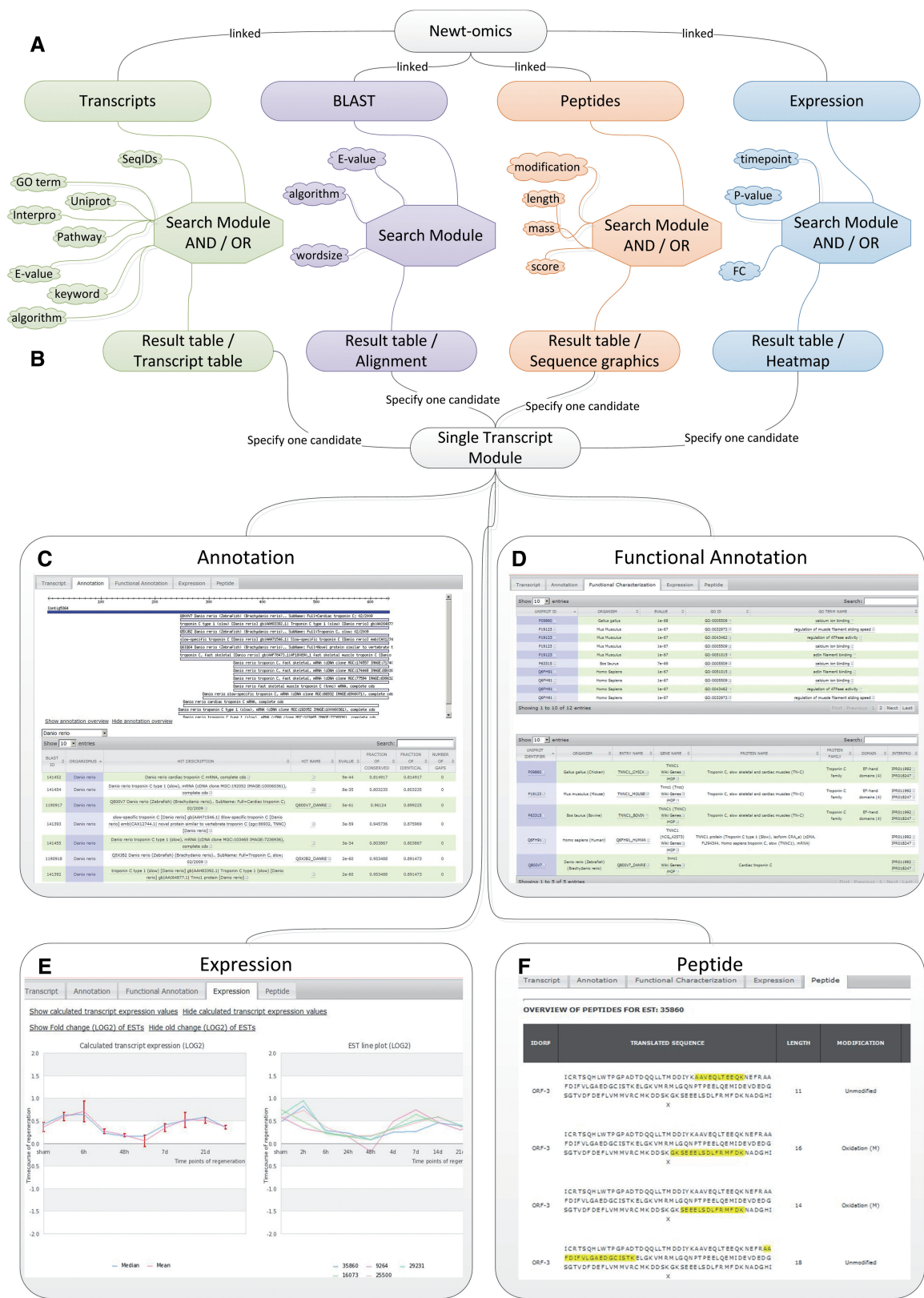
Figure 1. The hierarchical structure of the Newt-omics Graphical User Interface (GUI) (**A**) The four main windows Transcripts BLAST, Peptides, Expression allow searches for single transcripts or a group of transcripts from different affiliated data sources. (**B**) The results tables list transcripts that link to the 'Single Transcript Module'. All information linked to a selected transcript can be accessed via the four tabs Annotation (**C**), Functional Annotation (**D**), Expression (**E**) and Peptide (**F**).

links to the single 'Transcript' module (Figure 1B), which displays the cDNA sequence of a selected transcript and lists the IDs and coordinates of the individual ESTs that comprise the transcript. From this central object, all affiliated information can be accessed via the tabs, 'Annotation', 'Functional Annotation', 'Expression' and 'Peptide'.

The 'Annotation' view lists all information related to similarity searches for the selected transcript. The list can be sorted and selected by source organism, by hit description, corresponding identifiers from public databases, and for statistics of an alignment, like *e*-value, fraction of conserved, fraction of identical and number of gaps. Further information to a selected transcript is provided by an alignment graphics and an annotation overview that shows all organisms for which a similarity has been identified and displays the proportion of total annotations for each organism (Figure 1C). The 'Functional Annotation' view lists and links to GO terms affiliated to an individual transcript. This list can be sorted by organism, *e*-value, UNIPROT ID, GO identifier and GO term name. A submenu below contains more detailed information about identified protein domains or Interpro crosslinks to UNIPROT identifiers. This list can also be sorted by organisms, by UNIPROT ID and corresponding protein names and abbreviations. It also provides links to wikigenes (24), and iHOP (25), corresponding protein families and protein domains with associated interpro links (Figure 1D).

The 'Expression' view provides a detailed view of calculated expression values for a transcript and the individual ESTs, comprising a selected transcript. Graphs of calculated mean and median fold changes with standard deviation are displayed for a transcript on the left, whereas expression changes for individual ESTs comprising the transcript are displayed on the right. Calculated fold changes for the selected transcript and fold changes of the individual ESTs can be displayed in separate tables (Figure 1E).

The 'Peptide' view provides detailed information for experimentally validated peptides, corresponding to a selected transcript. Since peptide identifications are based on EST sequences, the matching peptides are highlighted within translated ESTs that constitute a transcript. Detailed information to a peptide including length, modification, modified sequence, Mascot score and mass to charge ratio is listed right to the corresponding peptide (Figure 1F).

DISCUSSION

Newt-omics database is the first integrated repository for the red spotted newt. It constitutes a reference resource for transcripts and proteins that are expressed in the newt heart. Furthermore, it allows a detailed gene expression analysis of newt heart regeneration. The database scheme was developed with a focus on (i) pre-processing of biological data and bioinformatics analysis, (ii) extensibility and (iii) modelling of biological data. Our approach features a user-friendly web interface, which allows

intuitive access for researchers with limited bioinformatical training.

For high content data sets that are based on large batch sequencing or microarray time course experiments, appropriate bioinformatics is crucial for the generation of useful annotations and functional assignments. Raw data are usually of limited value for non-bioinformaticians. Therefore, we decided to store only bioinformatically pre-processed data sets in Newt-omics. Similarly, we wanted to visualize the quality of our bioinformatics analysis by graphical representation, rather than giving statistical values, which facilitates access to non-specialists. For example, data plots visualize the uniformity of transcript expression during the time course of heart regeneration. Line plots for individual members of a transcript help to identify irregular EST expression patterns that might arise due to imperfect sequence assembly or by alternative splice isoforms. The graphical peptide view identifies positions of peptides within an ORF, that allows discrimination between coding and non-coding areas of a transcript. This feature is extremely helpful to analyse transcripts that lack any sequence homology, since the existence of matching peptides identify transcripts as true protein coding. Since we performed homology searches on multiple organisms and source databases, it is possible to sort identified sequences by source organism, significance of similarity and description. Such a sorting approach provides information about the degree of conservation between multiple species. Likewise, involvement of molecules in regenerative processes known from other organisms might be uncovered by their functional annotations. Newt-omics allows identification of GO terms of interest that are associated with tissue development or regeneration in mouse, human, fish and chicken.

Another important feature of our database is the ability to update each part of the database separately without disturbing the integrity of the database. It is evident that newly emerging sequencing technologies will make traditional Sanger-based sequencing approaches of new model organisms obsolete even if no 'matrix' for assembly sequence reads is available. Although recent algorithms for *de novo* assemblies of short reads are still challenging, the rapid development of this field will help to overcome existing obstacles. Moreover, the Sanger-sequencing reads stored in our database provide a solid basis for future assembly projects and facilitate gene expression experiments based on arrays and short read sequencing. Future updates of the database will include a more comprehensive set of transcriptome sequencing data derived from other tissues than regenerating heart. This data set will also allow to increase the number of peptides that can be identified from mass spectrometry measurements.

The design of our database emphasizes the linkage between the different data sets. The web-based frontend enables an intuitive 'from each view to any other view' within the database. We have focused primarily on the 'transcript' as the central object of the database, which might help to address biologically relevant processes. In contrast, the central element of other relational databases with a similar approach relies on genome information (26), which is more difficult to link to biological processes.

Taken together, our database provides molecular insights into a valuable, yet relatively little studied model organism that allows detailed characterization of regenerative processes. The database provides a comprehensive repository not only for researchers working with *N. viridescens* but also for others in closely related biological disciplines like developmental biology and regenerative medicine. Newt-omics complements large publicly available databases and provides detailed information for research in regenerative medicine ranging from salamanders to humans (27–29).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figure 1.

REFERENCES

1. Bowes,J.B., Snyder,K.A., Segerdell,E., Gibb,R., Jarabek,C., Noumen,E., Pollet,N. and Vize,P.D. (2008) Xenbase: a Xenopus biology and genomics resource. *Nucleic Acids Res.*, **36**, D761–D767.
2. Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Howe,D.G., Knight,J., Mani,P., Martin,R., Moxon,S.A. *et al.* (2011) ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res.*, **39**, D822–D829.
3. de la Cruz,N., Bromberg,S., Pasko,D., Shimoyama,M., Twigger,S., Chen,J., Chen,C.F., Fan,C., Foote,C., Gopinath,G.R. *et al.* (2005) The Rat Genome Database (RGD): developments towards a phenome database. *Nucleic Acids Res.*, **33**, D485–D491.
4. Diez-Roux,G., Banfi,S., Sultan,M., Geffers,L., Anand,S., Rozado,D., Magen,A., Canidio,E., Pagani,M., Peluso,I. *et al.* (2011) A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol.*, **9**, e1000582.
5. Drysdale,R. (2008) FlyBase: a database for the Drosophila research community. *Methods Mol. Biol.*, **420**, 45–59.
6. Brockes,J.P. (1997) Amphibian limb regeneration: rebuilding a complex structure. *Science*, **276**, 81–87.
7. Iten,L.E. and Bryant,S.V. (1976) Regeneration from different levels along the tail of the newt, Notophthalmus viridescens. *J. Exp. Zool.*, **196**, 293–306.
8. Tsonis,P.A. (1991) Amphibian limb regeneration. *In Vivo*, **5**, 541–550.
9. Call,M.K., Grogg,M.W. and Tsonis,P.A. (2005) Eye on regeneration. *Anat. Rec. B New Anat.*, **287**, 42–48.
10. Henry,J.J. and Tsonis,P.A. (2010) Molecular and cellular aspects of amphibian lens regeneration. *Prog. Retin. Eye Res.*, **29**, 543–555.
11. Berg,D.A., Kirkham,M., Beljajeva,A., Knapp,D., Habermann,B., Ryge,J., Tanaka,E.M. and Simon,A. (2010) Efficient regeneration by activation of neurogenesis in homeostatically quiescent regions of the adult vertebrate brain. *Development*, **137**, 4127–4134.
12. Parish,C.L., Beljajeva,A., Arenas,E. and Simon,A. (2007) Midbrain dopaminergic neurogenesis and behavioural recovery in a salamander lesion-induced regeneration model. *Development*, **134**, 2881–2887.
13. Bader,D. and Oberpriller,J.O. (1978) Repair and reorganization of minced cardiac muscle in the adult newt (Notophthalmus viridescens). *J. Morphol.*, **155**, 349–357.
14. Borchardt,T. and Braun,T. (2007) Cardiovascular regeneration in non-mammalian model systems: what are the differences between newts and man? *Thromb. Haemost.*, **98**, 311–318.
15. Oberpriller,J.O. and Oberpriller,J.C. (1974) Response of the adult newt ventricle to injury. *J. Exp. Zool.*, **187**, 249–253.
16. Brockes,J. and Kumar,A. (2005) Newts. *Curr. Biol.*, **15**, R42–44.
17. Borchardt,T., Looso,M., Bruckskotten,M., Weis,P., Kruse,J. and Braun,T. (2010) Analysis of newly established EST databases reveals similarities between heart regeneration in newt and fish. *BMC Genomics*, **11**, 4.
18. Looso,M., Borchardt,T., Kruger,M. and Braun,T. (2010) Advanced identification of proteins in uncharacterized proteomes by pulsed in vivo stable isotope labeling-based mass spectrometry. *Mol. Cell Proteomics*, **9**, 1157–1166.
19. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396–D403.
20. Kruger,M., Moser,M., Ussar,S., Thievessen,I., Luber,C.A., Forner,F., Schmidt,S., Zanivan,S., Fassler,R. and Mann,M. (2008) SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. *Cell*, **134**, 353–364.
21. Schwanhausser,B., Gossen,M., Dittmar,G. and Selbach,M. (2009) Global analysis of cellular protein translation by pulsed SILAC. *Proteomics*, **9**, 205–209.
22. Elias,J.E. and Gygi,S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods.*, **4**, 207–214.
23. Cheng,X.Y., Chen,X.H. and Hua,J. (2011) The Overview of Entity Relation Extraction Methods. *Comm Com Inf Sc*, **134**, 749–754.
24. Hoffmann,R. (2008) A wiki for the life sciences where authorship matters. *Nat. Genet.*, **40**, 1047–1051.
25. Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet.*, **36**, 664–664.
26. Zhou,P., Emmert,D. and Zhang,P. (2006) Using Chado to store genome annotation data. *Curr. Protoc. Bioinformatics*, Chapter 9, Unit 9.6.
27. Antos,C.L. and Tanaka,E.M. (2010) Vertebrates that regenerate as models for guiding stem cels. *Adv. Exp. Med. Biol.*, **695**, 184–214.
28. Ausoni,S. and Sartore,S. (2009) From fish to amphibians to mammals: in search of novel strategies to optimize cardiac regeneration. *J. Cell Biol.*, **184**, 357–364.
29. Menger,B., Vogt,P.M., Kuhbier,J.W. and Reimers,K. (2010) Applying amphibian limb regeneration to human wound healing: a review. *Ann Plast Surg*, **65**, 504–510.